

머신러닝을 이용한 경기도 화재위험요인 예측분석 Predictive Analysis of Fire Risk Factors in Gyeonggi-do Using Machine Learning

서민송¹⁾ · 에베르 엔리케 카스티요 오소리오²⁾ · 유환희³⁾
 Seo, Min Song · Castillo Osorio, Ever Enrique · Yoo, Hwan Hee

Abstract

The seriousness of fire is rising because fire causes enormous damage to property and human life. Therefore, this study aims to predict various risk factors affecting fire by fire type. The predictive analysis of fire factors was carried out targeting Gyeonggi-do, which has the highest number of fires in the country. For the analysis, using machine learning methods SVM (Support Vector Machine), RF (Random Forest), GBRT (Gradient Boosted Regression Tree) the accuracy of each model was presented with a high fit model through MAE (Mean Absolute Error) and RMSE (Root Mean Squared Error), and based on this, predictive analysis of fire factors in Gyeonggi-do was conducted. In addition, using machine learning methods such as SVM (Support Vector Machine), RF (Random Forest), and GBRT (Gradient Boosted Regression Tree), the accuracy of each model was presented with a high-fit model through MAE and RMSE. Predictive analysis of occurrence factors was achieved. Based on this, as a result of comparative analysis of three machine learning methods, the RF method showed a MAE = 1.765 and RMSE = 1.876, as well as the MAE and RMSE verification and test data were very similar with a difference between MAE = 0.046 and RMSE = 0.04 showing the best predictive results. The results of this study are expected to be used as useful data for fire safety management allowing decision makers to identify the sequence of dangers related to the factors affecting the occurrence of fire.

Keywords : Prediction of Fire Risk Factors, Multiple Regression Analysis, Machine Learning, Support Vector Machine, Random Forest, Gradient Boosted Regression Tree

초 록

화재는 막대한 재산과 인명피해를 초래하고 있으며 크고 작은 화재가 지속해서 발생하고 있다. 따라서 본 연구는 화재 유형별로 화재에 영향을 미치는 각종 위험요인을 예측하고자 한다. 전국에서 화재 발생 건수가 가장 많은 경기도를 대상으로 화재발생위험요인 예측분석을 실시하였다. 또한, 머신러닝 방법인 SVM, RF, GBRT를 활용하여 각 모형의 정확성을 MAE, RMSE를 통해 적합도가 높은 모형을 제시하였으며 이를 토대로 경기도 화재발생요인 예측분석을 실시하였다. 머신러닝 방법 3가지를 비교분석한 결과 RF가 MAE 1.517, RMSE 1.820으로 나타났으며 MAE, RMSE 검증데이터 및 시험데이터의 경우 MAE값 0.024, RMSE값 0.12의 차이로 매우 유사하게 나타나 가장 우수한 예측력으로 나타났다. RF기법을 적용하여 분석한 결과 공통적으로 발화장소가 화재발생에 가장 큰 영향을 주는 위험요인으로 나타났다. 이러한 연구 결과는 화재발생에 영향을 주는 요인들의 위험순서를 파악하여 화재안전관리의 유용한 자료로 활용될 것으로 예상된다.

핵심어 : 화재발생위험요인예측, 다중회귀분석, 머신러닝, SVM, RF, GBRT

Received 2021.07.28, Revised 2021.08.09, Accepted 2021.12.14

1) Member, Dept. of Urban Engineering, Gyeongsang National University (E-mail: msong7938@gmail.com)

2) Member, Dept. of Urban Engineering, Gyeongsang National University (E-mail: ever.castillo.osorio@gmail.com)

3) Corresponding Author, Member, ERI, Dept. of Urban Engineering, Gyeongsang National University (E-mail: hhyoo@gnu.ac.kr)

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. 배경 및 목적

크고 작은 화재는 매년 발생하고 있으며 화재의 위험성은 지속적으로 대두되고 있다. 화재의 경우 언제 어디서나 일어날 수 있는 재난 중 하나이므로 이를 예방하고 대책을 세우는 것이 중요하다. 2019년도 화재 통계자료에 따르면 화재 중 부주의가 가장 많이 발생하고 있으며 부주의로 발생한 화재의 경우 작은 화재부터 대형화재로 이어지기까지 피해 규모가 다양하다. 또한, 도시 내의 건축물은 더욱 고층화 및 심층화되어 다양한 용도를 지닌 복합 경향을 나타내고 있으며 이는 대형화재로 이어질 수 있어 화재 예방이 절실한 실정이다. 우리나라의 화재는 2018년 기준으로 42,338건이 발생하였고, 인명피해는 2,594명 재산피해는 5,597억 원으로 발생하였다. 또한, 2014년에서 2018년까지의 평균 화재발생 건수는 43,300건이며 인명피해는 2,218명, 재산피해는 4,651억 원이 발생하였다(National Fire Data System, 2019). 화재는 사회·경제·지리·자연환경 등 다양한 요인에 의해 발생하고 있다. 우리나라의 화재요인은 발화요인으로 나뉘게 되며 발화요인이란 발화열원에 의하여 발화로 이어진 연소현상에 영향을 준 인적·물적·자연적 요인을 말한다(Kim, 2014)

화재는 막대한 재산과 인명피해를 초래하기 때문에 이에 대한 대책마련을 위해 한국 소방청에서 2007년부터 국가화재정보센터를 개설하였다. NFDS (National Fire Data System)는 화재가 발생한 곳을 조사하여 얻은 정보를 데이터화하고 있으며, 2009년부터는 국민들이 홈페이지를 통해 지역별 화재 현황과 화재통계를 열람할 수 있도록 관리되고 있다. 그러나 NFDS의 경우 단순 통계만 제공되고 있어 일반적인 사항만 알 수 있을 뿐 심층적인 위험 정도는 알기 어렵다. 이러한 통계자료는 일선에서 활동하고 있는 소방관계자들에게 화재를 자주 발생시키는 요인에 대해 설명해 줄 수는 있어도 화재가 언제 어느 정도 발생하게 될지는 알려주지 못한다.

화재의 심각성이 대두되고 있음에 따라 화재 관련 연구는 활발히 진행 중이며 화재 발생과 관련한 연구 동향을 살펴보면 다음과 같다.

국내 연구 동향의 경우 Lee *et al.*(2010)는 국가화재정보시스템의 2007년부터 2009년까지 강원지역에서 발생한 화재발생 자료에 대해 통계분석 방법 가운데 하나인 의사결정분석(decision tree analysis)방법을 이용하여 화재유형 분석을 실시하였다. 요인변수들은 발화환경, 화재원인, 진화요인으로 분류하였고, 목적변수로는 화재피해와 화재건수를 선정하여 분석하였다. Shin *et al.*(2010)는 통계적 분석방법을 통해 수도권 지역(서울특별시, 인천광역시, 경기도)의 화재발생 유형

에 대한 분석을 실시하였다. 2007년과 2008년 화재발생 자료를 대상으로 교차분석과 빈도분석을 비롯하여 의사결정트리 분석 등을 실시하였다. Seo *et al.*(2020)는 진주시를 대상으로 10년간 화재데이터를 사용하여 토지용도별, 시설물 유형별 그리고 인문사회적 요인을 고려하여 화재 발생의 유의성을 분석하였으며 최종 도출된 요인들을 중심으로 공간가중치를 적용하여 OLS (Ordinary Least Square) regression analysis, SLM (Spatial Lag Model), SEM (Spatial Error Model) 세 모형을 비교 분석하였다. Hong *et al.*(2018)는 인공지능을 활용한 데이터 기반 화재분석 및 예측·대응시스템 개발에서 최근 빅데이터 기술을 활용한 데이터 상관관계 분석 및 인공지능 기술 활용, 패턴 인식기술로 분류 및 예측, 회귀기법을 이용 화재사고 데이터 기반의 예측 및 대응시스템을 구축방안을 제시하였다.

국의 연구 동향의 경우 Hastie *et al.*(2016)은 영국을 중심으로 2010년부터 2013년까지 주거지역의 화재발생 건수를 종속변수로 설정하고 사회경제적 및 인구통계학적 자료를 독립변수로 설정하여 회귀분석을 실시하였으며 65세 미만 1인 가고, 5년 이상 일을 하지 않은 사람 또는 일을 한적이 없는 사람, 흑인의 수에 따라 화재가 많이 발생한다는 결과를 도출했다. Lu *et al.*(2013)는 2002년부터 2009년까지 화재 중 3명 이상의 사망자가 나왔던 화재를 분석하여 빈도·피해 분포와 시간·크기의 상관관계를 규명하였다. David *et al.*(2010)는 호주 퀸즐랜드 남동부를 대상으로 주택가 화재에 대해 베이저안 방법론을 적용하여 공간 예측을 실시하였다. 그 결과 1년 동안의 예상 화재 건수는 연간 2건에서 25건의 화재가 발생하는 것으로 분석되었으며 이를 지도화하여 화재위험을 줄이기 위한 예방방안을 제시하였다. Rifan *et al.*(2019)은 호주 멜버른을 중심으로 10년간 화재데이터를 마르코프 연쇄 방법을 사용하여 공간 및 시간을 중심으로 모델의 매개변수를 설정하여 화재가 취약한 지역의 공간 패턴을 분석하였으며 소방기관의 위치와 화재 위험 교육 프로그램에 대한 방안을 제시하였다.

현재까지의 화재 관련 연구는 제한된 정보에 의한 단순한 통계분석에 기초하여 결정되고 있으며, 데이터의 부정확성, 미흡한 분석정보로 인하여 화재대응 및 대책에 효율적인 의사결정이 이루어지지 못하고 있는 실정이다.

따라서 본 연구는 화재 유형별로 화재에 영향을 미치는 각종 위험 요인(시기, 원인, 기상, 인적요인 등)예측을 실시하고자 한다. 소방청에서 제공한 3년간(2017년~2019년) 데이터를 시군구 단위를 중심으로 사용하였으며 전국에서 화재발생 건수가 가장 많은 경기도를 대상으로 화재발생요인 예측분석을 실시하였다. 또한, 경기도 화재 중 가장 큰 발생요인인 부주의

(13,295건), 전기적(7,517건), 기계적 요인(3,947건)을 중심으로 비교분석하였다. 분석을 위해 다중회귀분석(MRA)모형과 머신러닝 방법인 SVM, RF, GBRT를 활용하여 각 모형의 정확성을 MAE, RMSE를 통해 적합도가 높은 모형을 제시하였으며 이를 토대로 경기도 화재발생요인 예측분석을 실시하였으며 본 연구의 연구흐름도는 Fig. 1과 같다.

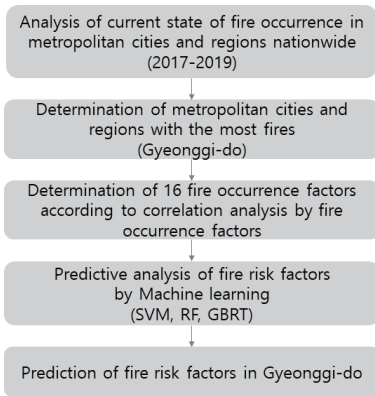


Fig. 1. Research flowchart

2. 머신러닝 이론

2.1 SVM (Support vector machine)

SVM 알고리즘은 Vapnik *et al.*(1963), Vapnik *et al.*(1964)에 의해 발전된 비선형 일반화 알고리즘으로서 통계적 학습 이론(statistical learning theory)의 견고한 기반이 되고 있다. 기본적인 SVM은 이진분류문제에 널리 이용되며, hyperplane을 중심으로 한쪽은 positive 클래스, 다른 한쪽은 negative 클래스로 나눈다(Hwang *et al.*, 2004). SVM의 가장 기본적인 아이디어는 두 개의 범주(positive, negative 클래스)를 구성하는 데이터들을 가장 잘 분리해 낼 수 있는 초평면(hyperplane)을 찾는 것이다. SVM에서는 저차원의 데이터를 고차원의 데이터 값으로 매핑(mapping)함에 따라 계산량 증가와 같은 문제점이 발생할 수 있는데 이는 커널(Kernel) 함수를 사용하여 해결가능하다. 커널 함수에는 가우시안 방사 기저 함수 커널(Gaussian Radial basis function kernel), 시그모이드 커널(sigmoid kernel), 선형 커널(linear kernel), 다항식 커널(polynominal kernel) 등이 있다. 이 중 어떤 커널을 사용할 것인지에 대해서는 합리적인 규칙이 정해져 있지 않으며 커널 함수별 성능도 큰 차이가 없기 때문에, 커널 함수의 결정은 데이터의 형태, 훈련 데이터의 총량, 속성 간의 관계를 고려한 다양한 시도와 평가를 기반으로 결정된다고 볼 수 있

다(Jang, 2018)

SVM은 분류와 회귀 모두 적용 가능하며 정확성이 높다. 또한 노이즈에 민감하지 않는 장점이 있다. 하지만 초모수 결정에 신중해야하고 입력데이터의 관찰치, 속성이 많아지면 학습시간이 오래 소요된다. 또한, 해석이 어렵고 결과값의 도출 이유를 확인하기 어려운 단점이 있다.

2.2 RF (Random forest)

RF는 Breiman *et al.*(2001)에 의해 제시된 앙상블 학습(Ensemble Learning) 모형으로 부트스트랩(bootstrap) 방식을 이용하여 다수의 표본을 생성하고 결정트리(Decision Tree)모형을 적용하여 그 결과를 종합하는 방법으로, 다수의 결정트리 모형을 결합시킨 형태이다(Seo, 2016).

랜덤 포레스트는 배깅과 같은 방법으로 부트스트랩된 훈련 자료를 이용하여 의사결정 나무를 만든다(Park, 2018)

하나의 트리로 구성된 결정트리 모형은 개념이 단순하고 시각화가 가능하기 때문에 해석이 쉽다는 장점이 있다. 그러나 예측치의 분산이 높기 때문에 예측 결과의 안정성이 떨어지며 다른 비선형 모형과 비교했을 때 추정 결과의 정확성이 낮은 편이라는 단점이 있다. 이러한 한계점은 단일 트리가 아닌 다수의 트리 결과를 종합하게 되면 정확도가 높아지는 경우가 많은데 이러한 방법을 앙상블 방법이라고 하며, 대표적으로 배깅(Bagging), RF, 부스팅(Boosting) 등이 있다.

단일 결정트리 모형이 분산이 높다는 문제점을 해결하기 위해서 데이터로부터 부트스트랩을 통해 B개의 데이터 집합(dataset)을 확보하고, B개의 회귀트리 결과를 각각 계산한 후, 마지막으로 이를 평균하여 최종 예측치를 결정할 수 있으며 이러한 방법을 배깅이라고 한다. 이는 Eq. (1)와 같으며 \hat{f}_{bag} 는 배깅방법을 활용한 최종 예측치를 의미하고 (f)는 예측모형을 나타내며 B 는 학습데이터의 총 개수, b 는 각 데이터셋을 의미한다.(Lee, 2015).

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}_b(x) \tag{3}$$

의사결정트리 모형보다 다수의 트리를 결합시킨 RF는 모형은 안정적이며 모형의 정확도가 현저히 개선된다. 또한, 대수의 법칙(Law of Large Numbers)에 의해 과적합을 피할 수 있으며(Breiman, 2001), 잡음이나 이상치에 영향을 크게 받지 않는다(Kim *et al.*, 2016). 다만, RF 적용에 있어서 부트스트랩 표본을 몇 개로 할 것인지, 각 마디에서 설명변수의 개수를 몇 개로 할 것인지, 결과 종합 시 선형 결합을 어떻게 할 것

인지는 여전히 연구자가 선택할 사항이다. 랜덤포레스트는 다중선형 회귀분석 및 로지스틱 회귀분석과 달리 개별 변수가 통계적으로 얼마나 유의한지에 대한 변수의 중요도를 측정할 수 있다. 변수의 중요도를 구하는 과정은 총 세 번의 단계를 거치게 되며 첫 번째로 데이터 집합에 대해 에러를 구하고 두 번째로 특정 변수의 값을 임의로 뒤섞은 데이터 집합에 대해 에러를 구한다. 마지막 세 번째 개별 변수의 중요도는 두 번째와 첫 번째 에러의 평균과 분산은 고려하여 추정하게 된다. 이를 사이킷런 혼련이라고 하며 세 번의 단계를 거치면 최종적으로 각 변수별로 정규화한 중요도 값을 얻게 된다. 중요도 점수는 모든 트리들에 대해서 원본 데이터 집합의 오차와 값이 치환된 데이터 집합의 오차의 차이 평균으로 정의된다. 랜덤포레스트의 경우 분류와 회귀 모두 적용 가능하며 정확성이 높다. 또한, 노이즈, 이상치에 민감하지 않고 과적합을 피할 수가 있는 장점이 있다. 반면 초모수 결정에 신중해야 하는 점과 속도와 메모리의 상대적 이용이 크다는 단점이 있다.

2.3 GBRT (Gradient boosted regression tree)

GBM은 높은 유연성으로 인해 특정 데이터 중심 작업에 맞게 사용자 지정이 가능하다. 모델설계에 많은 자유를 부여하므로 시행착오의 문제에서 가장 적합한 손실 함수를 선택하게 한다. 또한, GBM 구현하기가 비교적 간단하여 다양한 문제뿐만 아니라 지반공학분야에서도 상당한 성공을 거두었다. GBM은 전형적으로 기초 학습자로서 CART (Classification And Regression Tree), 결정나무를 사용하며 이를 GBRT로 불린다. GBRT는 RF와 마찬가지로 분류와 회귀 모두 적용이 가능하며 정확성이 높고 스케일 조정이 필요 없다. 또한, 연속적인 특성에도 잘 작동되는 장점이 있다. 하지만 고차원 데이터는 잘 작동하지 않는 단점이 있다(Lee, 2020)

2.4 머신러닝기법을 이용한 예측력 평가

머신러닝을 이용한 화재발생 예측분석에 있어서 알고리즘 탐색 및 모델링에서는 실무적 운용을 위한 알고리즘을 선택하고 해당 알고리즘을 모델링하는 방법에 대한 결정이 이루어지는데, 알고리즘과 모델링 방법의 선택 기준은 '활용 가능성(applicability)'이다. 따라서 '활용 가능성'을 어떻게 판단할 것인지에 대한 검토가 필요하다(Bae, 2001).

화재 위험도 예측에 있어서 머신러닝의 활용을 위해서는 머신러닝 모델의 정확성이 인정되어야 한다. 모형별 MAE와 RMSE 비교는 수치예측모델의 정확성 평가에 자주 사용되는 방법이며, 머신러닝 모델의 정확성 평가에도 사용이 가능하다. MAE는 Eq. (2)와 같이 측정 값(\hat{y}_i)과 실제 값(y_i)과

의 절대 오차의 평균을 계산한 값이며, RMSE는 Eq. (3)와 같이 측정 값(\hat{y}_i)과 실제 값(y_i)의 오차의 평균 제곱근을 계산한 값이다.

$$MAE = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y_i| \quad (2)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2} \quad (3)$$

머신 러닝의 정확도를 높이기 위해서는 데이터 분포 설명이 가능한 전역 최적해(global optimum)를 도출해야 한다. 전역 최적해를 도출하기 위해서는 각각의 머신 러닝 방법들 중에서 초모수(hyperparameter)를 최적화하는 과정이 필요하다. 머신 러닝의 초모수를 선택할 수 있는 기준이나 방법은 정립되어 있지 않으며, 머신 러닝의 실제 데이터에 맞춰 초모수를 변화시키면서 오차가 최소화되는 초모수를 찾는 것이 일반적이다. 즉, 여러번의 시행착오 속에 최적의 초모수를 도출할 수 있다. 최적의 초모수 도출은 혼련데이터를 이용한 학습 과정에서 이루어지게 되는데, 이 과정은 k-겹 교차검증 방법이 사용된다. 검증 과정은 k번 반복하면서 일반화 오차가 가장 낮은 초모수를 최종모형으로 결정하게 된다.

초모수 결정에서의 가장 중요한 부분은 일반화 오차의 최소화 및 과적합 방지를 위한 노력이 필요하다. 따라서, 본 연구에서는 초모수를 변화시키면서 10겹 교차검증을 실시하였으며, 초모수의 탐색 범위를 설정하고 초모수가 변화됨에 따른 검증 데이터(validation data)의 MAE 및 RMSE가 가장 낮은 초모수를 시험 데이터를 적용할 모형별 초모수를 결정하였다. 다만, 초모수 변화에 따른 MAE 및 RMSE의 차이가 크지 않고 과적합이 발생할 수 있다는 점을 고려하여 MAE 및 RMSE가 최소화되는 모형 외에 예비적 모형을 추가로 선정하였으며, 이에 따라 본 주제에서 시험 데이터를 적용할 모형은 각 머신 러닝 방법별로 2개씩 선정되었다(Bae, 2001).

전체 화재발생데이터 중 혼련데이터, 시험데이터로 구분하고 Python 및 R Studio 프로그램을 통해 분석을 실시하였으며 비율은 혼련데이터 70%, 시험데이터 30%로 분석하였다. 또한, 모형 간의 예측 정확성을 비교분석 하기 때문에 모형 모두 동일한 학습데이터를 활용하였다.

3. 결과분석

3.1 분석자료

본 연구는 소방청에서 제공한 2017년부터 2019년까지 3년

간의 전국화재 데이터를 사용하였다. 또한, 화재발생요인 예측분석을 위해 소방청에서 제공한 화재발생요인을 활용하였으며 2017년도 56개, 2018년도 22개, 2019년 22개의 요인들로 구성되어 있다. 소방청에서 제공하는 정보는 화재가 발생한 곳의 화재 시간정보, 화재유형, 화재원인 등에 의한 정보로만 구성되어 있다. 그러나 화재는 여러 가지 다양한 요인에 의해 발생되고 인적요인 또는 사회적 요인에 의해 화재발생이 일어나기도 한다. 따라서 본 연구는 화재조사 데이터 이외의 인문·사회적 요인인 시군구 면적, 인구밀도, 유동인구를 추가하여 25개의 데이터로 설정하였으며 데이터는 소방청, 국가통계포털, 공공데이터포털, 금융빅데이터플랫폼에서 취득하였으며 사용된 변수는 다음과 같다. 화재발생확률 및 피해규모에 측 최적화에 적용할 수 있는 화재 시간 정보, 날씨정보, 대응 정보, 피해정도 등을 기준으로 변수를 선정하였으며, 25가지 변수는 화재발생일시(월/시/분/요일), 날씨(풍속/온도/습도/풍향), 출동소요시간, 시군구별 인구, 읍면동별 인구, 화재유형, 화재발생층, 화재진압시간, 사망자, 부상자, 인명피해, 재산피

해, 발화장소, 소방동원명수, 소방서거리, 안전센터거리로 이루어져 있으며 인문·사회적 요인은 시군구 면적, 인구밀도(인구/면적), 유동인구로 이루어져 있으며 Table 1.과 같다.

전국화재 발생 건수는 2017년도 44,169건, 2018년도 41,851건, 2019년도 42,338건으로 나타났으며 2018년도 대비 2019년도 487건으로 다시 증가하였다. 광역시도 단위별 3년간의 화재 발생 건수를 살펴보면 경기도가 29,060건으로 가장 많이 발생하였으며 다음으로 서울 18,227건, 경상남도 11,079건, 전라남도 8,233건, 경상북도 8,188건, 충청남도 7,985건, 부산광역시, 7,551건, 강원도 6,820건, 전라북도 6,061건, 인천광역시 4,848건, 대구광역시 4,492건, 충청북도 4,382건, 대전광역시 3,247건, 울산광역시 2,733건, 광주광역시 2,643건, 제주특별자치도 2,021건, 세종특별자치시 788건 순으로 발생하였으며 Fig. 2와 같다. 본 연구는 전국 광역시도 단위의 3년간 전국화재 발생 건수를 고려하여 전국에서 화재 발생 건수가 가장 높은 경기도(29,060건)를 대상으로 화재발생위험요인 예측분석을 실시하였다.

Table 1. Fire occurrence factor variables

Variables	Explanation
Month	Month of fire
Hour	Hour of fire
Minute	Minutes of fire
Day of the week	Day of the fire
Wind speed	Wind speed at that fire day
Temperature	Temperature at that fire day
Humidity	Humidity at that fire day
Wind direction	Wind direction at that fire day
Population by city, county and district	Population by city, county and district
Population by administrative	Population by Eup, Myeon and Dong
Fire type	Construction, dangerous Goods, automobiles, garbage Fire, etc.
Fire floor	Number of floors on fire in the building
Fire suppression time	Fire suppression time
Dead	Person deaded by fire
Injured	Person injured by fire
Human damage	Human damage
Property damage	Property damage
Ignition location	Facility
Number of firefighters	Number of firefighters mobilized at the time of dispatch
Fire station distance	Fire station distance from the source of the fire
Safety center distance	Safety center distance from the fire source
Departure time	Time required for dispatch from the fire site
Area	City, county, and gu area
Population density	Population density
Floating population	Floating population

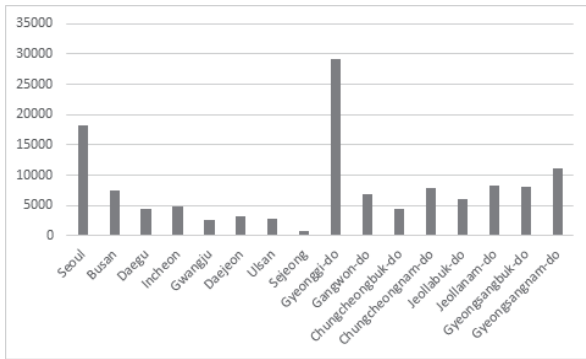


Fig. 2. Number of fires by metropolitan city and province nationwide(2017-2019)

3.2 화재 발생 요인 상관관계 분석

본 연구에서는 경기도 화재발생예측분석을 실시하기 위해 먼저 요인들 중 중요한 변인들을 찾아내기 위해 상관분석과 반복적인 회귀분석을 실시하였다. 또한, 다중공선성의 문제를 유발하는 요인들을 제외하였다. 이를 통해 유효한 변인들을 찾아냈으며 다중공선성의 문제가 되는 요인을 제외하고 분석을 진행하였다.

먼저 화재 발생과 각 요인과의 상관관계를 파악하였다. 상관관계 분석결과 상관성이 높은 변인들이 많이 분석되었으

며 다중공선성 발생 소지가 있다. 이러한 문제를 해결하기 위해 단계적 회귀분석을 실시하였다. 그 결과 9가지 요인들(분, 요일, 풍속, 풍향, 시군구별 인구, 읍면동별 인구, 화재진압시간, 사망자, 시군구 면적)은 제외하였으며 이러한 과정을 통해 최종 선정된 16가지 요인들(월, 시, 온도, 습도, 출동소요시간, 화재유형, 화재발생층, 부상자, 인명피해, 재산피해, 발화장소, 소방동원명수, 소방서거리, 안전센터거리, 인구밀도, 유동인구)을 대상으로 다중회귀분석을 실시하였다 (Table 2).

회귀분석 결과 모델의 설명력을 확인할 수 있는 R^2 가 0.8152로 나타났으며 81.5%의 설명력을 갖는 것으로 분석되었다. 또한, 유의수준 $\alpha = 0.05$ 에 모두 기각하고 있으므로 독립변수의 회귀계수는 종속변수와 독립변수가 서로 무관하다는 귀무가설이 발생하기 때문에 이들 모형에 포함된 독립변수들이 종속변수의 변량을 잘 설명해주고 있다.

3.3 화재요인예측을 위한 머신러닝 기법의 예측력 평가

본 연구에서는 경기도 화재발생위험요인예측분석을 실시하기 기계학습 방법인 SVM; RF, GBRT의 정확성을 MAE와 RMSE를 통해 비교분석하였으며 네 가지 방법을 이용해 경기도 전체 자료에 대해 분석을 실시한 후 예측력을 높은 모형을 채택하였다. 또한, 본 연구는 예측력이 높은 모형을 통

Table 2. Analysis of multiple regression

Model	Non-standardization factor		t	Standardization factor
	B	Standard error		
Month	.148	.011	2.326	.000
Time	-.751	.076	-3.116	.000
Temperature	.354	.245	16.882	.000
Humidity	.618	.071	9.332	.001
Departure time	.084	.098	7.485	.000
Fire floor	.035	.357	26.275	.002
Injured	-.197	.481	-2.423	.001
Casualties	.249	.687	6.485	.000
Property damage	.186	.012	15.524	.000
Ignition space	.684	.018	5.454	.001
Number of firefighters	.234	.082	12.549	.001
Fire station distance	.894	.723	4.987	.000
Safety center distance	.076	.059	3.481	.000
Fire type	-.741	.357	-5.251	.000
Population density	.443	.041	21.540	.000
Floating population	.017	.057	19.546	.000
R^2				0.8152

해 화재발생에 영향을 미치는 요인들을 분석하기 위해 화재 발생의 원인이 되는 발화요인을 중심으로 화재발생위험요인을 분석하였다. 발화요인에는 부주의, 전기적 요인, 화학적 요인, 자연적인 요인, 방화의심, 방화, 가스누출(폭발), 교통사고, 기계적 요인, 기타, 미상이 있다. 경기도의 발화요인별 건수는 Fig. 3과 같다.

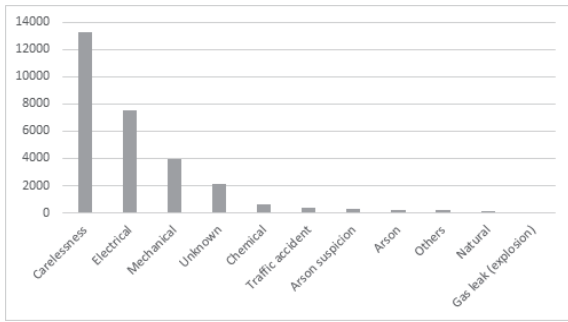


Fig. 3. Number of cases by fire occurrence factors in Gyeonggi-do

이를 토대로 경기도에서 발생한 화재들 중 가장 많이 발생한 부주의(13,295건), 전기적(7,517건), 기계적 요인(3,947건)을 중심으로 화재발생요인 예측분석을 실시하여 비교분석하였다.

3.3.1 SVM

SVM의 최적모형 결정을 위해서는 최적화를 적용할 커널함수(Kernel function), 오류에 대한벌칙을 제어하는 초모수(Hyper-parameter)인 C , 훈련데이터의 영향도 및 영향력 범위와 관련된 r , 그리고 훈련데이터 허용 에러율 ϵ 에 대한 결정이 필요하다. 기본적으로 커널함수를 적용하여 C 는 1~5, r 는 값에 따라 초평면의 곡률이 달라지게 되며 r 의 증가는 훈련 데이터에 대한 설명력은 증가하지만 과적합의 문제가 있을 수 있으므로 r 이 작은 모형을 최종 모형으로 선정해야한다. r 훈련 데이터의 영향력과 관련되어 있으므로 분석을 통해 0.2로 결정하였다. ϵ 는 0.01 또는 0.02로 결정하였다. 초모수의 결과는 최종적으로 MAE 및 RMSE의 값이 최소가 되는 모형을 SVM의 최적모형으로 결정하였다. 적합결과를 보면 MAE(4.309)는 $C=4, r=0.2, \epsilon=0.02$ 인 경우, RMSE(4.569)는 $C=4, r=0.2, \epsilon=0.02$ 인 경우 가장 낮은 것으로 나타나 최종 초모수로 결정하였다. 이러한 탐색 절차에 따른 경기도 SVM 적합결과는 각각 Table 3과 같다.

Table 3. Analysis of SVM

Parameter			Output	Error	
C	r	ϵ		MAE	RMSE
1	0.2	0.01	Output1	5.487	5.841
1	0.2	0.02	Output2	5.189	5.514
2	0.2	0.01	Output1	5.185	5.530
2	0.2	0.02	Output2	5.141	5.134
3	0.2	0.01	Output1	5.652	5.843
3	0.2	0.02	Output2	5.531	5.712
4	0.2	0.01	Output1	4.314	4.614
4	0.2	0.02	Output2	4.309	4.569
5	0.2	0.01	Output1	4.741	4.670
5	0.2	0.02	Output2	4.864	4.751

3.3.2 RF

RF는 초모수인 트리수를 50단위로 변화시키면서 k겹 교차검증을 실시한다. MAE값 및 RMSE값이 검증데이터 중에서 최소가 되는 트리수의 조합을 최종 모형으로 결정하였다. 트리수는 50부터 500까지이며 50간격의 단위로 분석을 실시하였다. 분석결과 트리수가 150을 초과하게 되면 MAE 및 RMSE는 각각 200 및 350 수준에서 큰 변화가 없으며, 트리수가 400인 경우 MAE값 및 RMSE값이 최소가 된다. 트리수가 150을 넘어서면 MAE(2.462) 및 RMSE(2.456)의 큰 변화가 없다는 점에서 트리수 200인 모형과 MAE(2.462) 및 RMSE(2.456)가 최소가 되는 트리수 400인 모형을 최종 모형으로 선택하였다. MAE 및 RMSE 개선여부를 고려하여 최종모형을 결정하였으며, Table 4는 경기도 RF 적합결과이다.

Table 4. Analysis of RF

Tree-number	Error	
	MAE	RMSE
50	3.154	3.248
100	3.355	3.324
150	2.861	2.786
200	2.462	2.456
250	2.154	2.435
300	2.165	2.340
350	2.124	2.378
400	1.541	1.832
450	1.751	1.176
500	2.154	2.785

3.3.3 GBRT

GBRT는 트리수와 학습률을 기준으로 주요 초모수를 결정한다. 학습률은 이진 트리의 오차(error)를 보정의 크기를 제어하는 초모수이다. 학습률은 0.1로 결정하였고, 트리수는 1000을 상한으로 설정하고 50부터 시작하여 MAE 및 RMSE 개선여부를 고려하여 최종모형을 결정하였으며, 경기도의 학습데이터를 이용한 적합 결과 다음과 같다.

분석결과 트리수가 600을 초과하게 되면 MAE 및 RMSE의 감소폭이 1이하로 축소되는 것으로 나타났으며, 트리수가 1000인 경우 MAE 및 RMSE는 최저가 되었다. 이에 따라 MAE(3.021) 및 RMSE(3.672)의 감소폭이 극도로 낮아지는 트리수가 400인 모형과 MAE(2.451) 및 RMSE(2.973)가 최저가 되는 트리수 1000인 모형을 최종 모형으로 선택하였다(Table 5).

Table 5. Analysis of GBRT

Tree-number	Error	
	MAE	RMSE
50	4.542	4.874
100	4.541	4.842
150	4.514	4.765
200	4.541	4.981
250	4.651	4.720
300	4.842	4.941
350	4.241	4.549
400	3.021	3.672
450	3.841	3.949
500	3.838	3.941
600	3.813	3.910
700	3.782	3.815
800	3.742	3.897
900	3.695	3.803
1000	2.451	2.973

3.3.4 머신러닝 기법별 예측력 비교분석

k겹 교차검증을 활용하여 검증데이터 및 시험데이터의 MAE값 RMSE값을 도출하였으며 값이 가장 작은 최종모형을 선정하였다(Table 6). 분석결과를 살펴보면 검증데이터의 MAE 및 RMSE보다 시험데이터의 MAE 및 RMSE 값이 더 낮은 것을 확인할 수 있으며 이는 시험데이터의 예측력이 우수하다는 것을 의미한다. 또한, 기존의 통계분석기법인 다중 회귀분석(MRA)의 MAE 및 RESE의 경우 머신러닝 기법들의 결과보다 정확성이 떨어지는 것을 확인할 수 있다. 따라서

머신러닝 기법 SVM, RF, GMRT를 기준으로 MAE 및 RMSE를 통해 비교분석한 결과 RF모형이 가장 예측력이 높은 것으로 분석되었으며 다음으로 GMRT, SVM 순으로 정확성이 높은 것으로 나타났다. 예측력이 가장 높은 RF모형을 기준으로 살펴보면 트리수가 200일 때 보다 400일 때가 MAE 및 RMSE가 가장 적은 수치로 나타났으며 검증데이터는 MAE 값 1.541, RMSE 값 1.832로 분석되며 시험데이터의 경우 MAE값 1.517, RMSE값 1.820으로 나타났다. 검증데이터 및 시험데이터 간의 MAE 및 RMSE의 차이는 SVM이 가장 큰 것으로 분석된다. 따라서 과적합 정도는 상대적으로 SVM이 높은 것으로 분석할 수 있다. 반면 RF분석결과 트리수 400을 기준으로 검증데이터 및 시험데이터의 MAE, RMSE의 경우 MAE값 0.024, RMSE값 0.12로 매우 유사하여 과적합의 정도가 상대적으로 낮은 것으로 분석된다.

Table 6. Comparative analysis of predictive power by model in Gyeonggi-do

		Validation data			Test data	
		MAE	RMSE		MAE	RMSE
MRA	-	-	-		120.061	156.870
SVM	$C=4, r=0.2, \epsilon=0.01$	4.314	4.614	output	4.106	4.147
	$C=4, r=0.2, \epsilon=0.02$	4.309	4.569		4.132	4.273
RF	estimators =200	2.462	2.456		2.439	2.421
	estimators =400	1.541	1.832		1.517	1.820
GMRT	estimators =400	3.021	3.672		2.981	3.513
	estimators =1000	2.451	2.973		2.325	2.719

3.3.5 화재요인별 RF기법의 분석결과

경기도 전체를 대상으로 머신러닝 분석을 실시한 결과 RF가 가장 유의미한 모형으로 나타났다. 따라서, 이를 토대로 경기도에서 발생한 화재들 중 가장 많이 발생한 부주의(13,295건), 전기적(7,517건), 기계적 요인(3,947건)을 중심으로 화재 발생요인 예측분석을 실시하여 비교분석하였다.

경기도의 화재발생 중 화재발생요인별(부주의, 전기적, 기계적) RF분석 결과는 Table 7과 같다. 부주의 요인의 RF 분석결과 트리수가 100을 초과하게 되면 MAE 및 RMSE는 각각 150및 350 수준에서 큰 변화가 없으며, 트리수가 400인 경우 MAE값 및 RMSE값이 가장 최소가 된다. 트리수가 100을 넘어서면 MAE 및 RMSE의 큰 변화가 없다는 점에서 트리수 100인 모형과 MAE값 및 RMSE값이 가장 최소가 되는 트리수 400인 모형을 최종모형으로 선택하였다.

전기적 요인의 RF 분석결과 트리수가 200을 초과하게 되면 MAE 및 RMSE는 각각 250 및 450 수준에서 큰 변화가 없으며, 트리수가 500인 경우 MAE값 및 RMSE값이 가장 최소가 된다. 트리수가 200을 넘어서면 MAE 및 RMSE의 큰 변화가 없다는 점에서 트리수 200인 모형과 MAE값 및 RMSE값이 가장 최소가 되는 트리수 500인 모형을 최종모형으로 선택하였다.

기계적 요인의 RF 분석결과 트리수가 50을 초과하게 되면 MAE 및 RMSE는 각각 100 및 250 수준에서 큰 변화가 없으며, 트리수가 300인 경우 MAE값 및 RMSE값이 가장 최소값이 된다. 트리수가 50을 넘어서면 MAE 및 RMSE의 큰 변화가 없다는 점에서 트리수 50인 모형과 MAE값 및 RMSE값이 가장 최소가 되는 트리수 300인 모형을 최종모형으로 선택하였다.

Table 7. Results of RF analysis by factors

Factor Tree-number	Carelessness		Electrical		Mechanical	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
50	3.512	3.811	3.540	3.787	3.985	4.028
100	2.851	2.948	3.421	3.840	3.812	3.942
150	2.737	2.946	3.143	3.482	3.807	3.910
200	2.702	2.938	2.846	3.055	3.803	3.921
250	2.614	2.928	2.810	2.923	3.751	3.830
300	2.619	2.911	2.706	2.851	2.782	2.812
350	2.507	2.884	2.652	2.829	2.799	2.820
400	1.912	2.043	2.569	2.731	2.913	3.069
450	2.016	2.215	2.432	2.607	3.018	3.198
500	2.230	2.348	1.814	1.968	3.284	3.296

예측력이 가장 높다고 판단된 RF를 기준으로 요인들 간의 상대적인 중요도를 분석하였다. 이는 향후 화재 발생에 영향을 줄 수 있는 요인들을 분석할 수 있다. Table 8은 RF를 기준으로 경기도의 훈련데이터를 활용하여 변수의 상대적 중요도를 측정한 것이며 이에 대한 결과는 다음과 같다. 부주의 요인을 기준으로 화재발위험요인을 예측한 결과 발화장소가 화재 발생에 가장 큰 영향을 미치는 것으로 나왔으며 다음으로 발생 시, 온도, 화재유형, 발생 월 등의 순으로 나타났다.

전기적 요인을 중심으로 분석한 결과 화재에 가장 큰 영향을 주는 요인은 발화장소로 나타났으며 다음으로 화재유형, 부상자, 출동소요시간, 재산피해 등의 순으로 나타났다.

기계적 요인을 기준으로 분석한 결과 발화장소가 향후 화재 발생에 가장 큰 영향을 미치는 것으로 나왔으며 화재유형, 소방

서로부터의 거리, 재산피해, 인명피해 등의 순으로 분석되었다.

Table 8. Evaluation of importance by RF-oriented factors

Factors Variables	Carelessness		Electrical		Mechanical	
	50	300	50	400	50	400
Month	10.14	8.51	1.27	1.74	0.14	0.15
Time	22.47	22.87	0.26	0.34	1.59	1.57
Temperature	0.74	0.87	0.41	0.57	0.08	0.08
Humidity	0.24	0.25	0.08	0.08	0.24	0.21
Departure time	0.34	0.94	4.14	4.28	0.20	0.21
Fire floor	0.41	0.54	0.09	0.09	0.04	0.06
Injured	9.14	9.81	9.14	9.57	6.78	7.85
Casualties	0.45	0.65	0.57	0.60	0.15	0.74
Property damage	0.10	0.21	2.24	2.37	10.26	10.34
Ignition space	25.74	25.75	34.45	35.78	25.78	26.47
Number of firefighters	0.41	0.49	0.45	0.15	0.32	0.35
Fire station distance	0.12	0.23	0.24	0.37	0.58	0.46
Safety center distance	0.34	0.18	0.19	0.10	12.26	13.34
Fire type	9.14	8.81	25.41	26.27	9.47	9.51
Population density	0.14	0.25	0.01	0.02	0.09	0.09
Floating population	5.41	5.78	0.65	0.82	0.08	0.08

3.3.6 화재위험요인별 중요도 종합평가분석

경기도를 대상으로 머신러닝 네 가지 방법을 사용하여 예측력이 가장 높은 RF모형을 중심으로 향후 화재발생 시 가장 많은 영향을 주는 요인을 중요도 기준 예측 분석하였다. 또한, 화재발생과 밀접한 관계가 있는 발화요인 대분류 중 발생 건수가 가장 많은 부주의, 전기, 기계를 기준으로 화재발생예측분석을 실시하였으며 RF 분석결과 중요도 순서는 Table 9와 같다.

Table 9. Evaluation of importance by RF-oriented factors

Factor	1	2	3	4	5
Carelessness	Ignition space (facility)	Time	Temperature	Fire type	Month
Electric	Ignition space (facility)	Fire type	Casualties	Departure time	Property damage
Machine	Ignition space (facility)	Fire type	Fire station distance	Property damage	Injured

중요도 순서 기준 요인별 분석을 실시하고자 소방청에서 제공 받은 데이터를 중심으로 요인별 종합평가를 실시하였다. 먼저 부주의 13,295건을 대상으로 중요도 순 화재발생요인을 확인해본 결과 발화장소가 화재 발생에 가장 많은 영향을 주는 위험요인으로 나타났으며 다음으로 발생 시,온도, 화재유형, 발생 월 순으로 나타났다.

전기적 요인 7,517건을 대상으로 중요도 순 화재발생요인을 확인해본 결과 발화장소가 화재 발생에 가장 많은 영향을 주는 위험요인으로 나타났으며 다음으로 화재유형, 부상자, 출동소요시간, 재산피해 순으로 향후 화재발생에 영향을 주는 요인으로 나타났다.

기계적 요인 3,947건을 대상으로 중요도 순 화재발생요인을 확인해본 결과 발화장소가 화재 발생에 가장 많은 영향을 주는 위험요인으로 나타났으며 다음으로 화재유형, 소방서로부터 거리, 재산피해, 인명피해 순으로 나타났다.

부주의, 전기적, 기계적 요인 모두 발화장소가 향후 화재가 발생할 수 있는 위험요인으로 예측되었으며 이에 대해 분석하면 다음과 같다.

먼저 부주의 요인의 경우 발화장소 중에서 공동주택이 3,446건으로 가장 높았으며 다음으로 단독주택 3,139건 공장시설 2,950건 음식점 1,767건으로 나타났다. 화재발생시간을 기준으로 보았을 때 19시-21시에 화재발생이 가장 많이 발생하며 다음으로 14시-15시에 화재발생 건수가 높은 것으로 분석된다. 온도의 경우 10도 이상 20도 미만일 때 화재 발생률이 가장 높은 것으로 나타났다. 또한, 전기적 요인의 경우 공동주택 969건으로 화재가 가장 많이 발생하였으며 다음으로 단독주택 891건, 공장시설 801건, 음식점 519건, 창고시설 387건 순으로 나타났다. 화재유형으로 보았을 때 건축, 구조물에 의한 화재가 가장 많이 발생하였으며 그 중 철근콘크리트에 의한 화재가 942건으로 가장 높은 것으로 나타났다. 기계적 요인의 경우 공장시설이 812건으로 가장 많이 나타났으며 다음으로 단독주택 326건, 공동주택 284건, 음식점 236건, 판매시설 98건 순으로 화재발생건수가 높은 것으로 분석된다. 또한, 화재 유형의 경우 건축, 구조물 중 철근콘크리트에 의한 화재가 가장 많이 발생한 것으로 나타났다.

4. 결론

본 연구는 3년간(2017년~2019년) 화재데이터를 취득하여 경기도 화재발생예측분석을 실시하기 위해 분석을 위해 다중회귀분석(MRA)모형과 머신러닝 방법인 SVM, RF, GBRT을 활용하여 각 모형의 정확성을 MAE, RMSE를 통해 적합

도가 높은 모형을 제시하였으며 이를 토대로 경기도에서 발생한 화재들 중 가장 많이 발생한 부주의, 전기적, 기계적 요인을 중심으로 화재발생위험 요인 예측분석실시하였으며 결과는 다음과 같다.

첫째, 전국화재 발생 건수를 분석한 결과 경기도가 28,852건으로 가장 높게 나타났으며 화재 발생 건수가 가장 높은 경기도를 대상으로 분석을 실시하였다. 경기 지역의 발화요인을 분석한 결과 경기 지역의 발화요인별/월별 발생확률로 가장 많은 화재발생 부분을 차지하는 발화요인은 부주의, 전기적, 기계적 요인으로 나타났다. 또한, 다중회귀분석에 따라 화재발생과 상관성이 있고 다중공선성 문제가 없는 16개의 화재발생요인은 월, 시, 온도, 습도, 출동소요시간, 화재발생층, 화재유형, 부상자, 인명피해, 재산피해, 발화장소, 소방동원명수, 소방서거리, 안전센터거리, 인구밀도, 유동인구로 분석되었다.

둘째, 경기 지역 전체를 중심으로 훈련 데이터 및 검증 데이터를 이용한 k겹 교차검증에 따라 결정된 최종모형을 이용하여 SVM, RF, GBRT의 평가지수(MAE, RMSE)에 따른 예측력을 비교분석한 결과 RF, GBRT, SVM 순으로 나타났다.

셋째, 머신러닝 모형 기법 중 가장 우수한 예측력을 나타낸 RF모형을 이용하여 화재 발생과 밀접한 관계가 있는 발화요인(대분류) 중 발생 건수가 가장 많은 부주의, 전기적, 기계적 요인을 기준으로 화재발생예측분석을 실시하였다. 세 가지 요인 중 화재 발생에 가장 큰 위험요인은 발화장소로서 부주의 및 전기적 요인의 경우 공동주택에서 화재발생이 가장 많이 나타났으며 기계적 요인의 경우 공장시설에서 화재가 가장 많이 발생하였다.

따라서 본 연구는 경기도의 3년간 화재 발생 데이터를 중심으로 머신러닝을 이용하여 요인별 화재발생예측을 실시함으로써 화재발생 안전관리대책마련에 유용한 자료로 활용될 것으로 기대된다.

감사의 글

본 연구는 과학기술정보통신부 이공학개인지초연구지원사업 기본연구(2021R1F1A106422811)의 연구비 지원에 의해 수행되었습니다.

References

Brieman, L. (2001), Random Forests, *Machine learning*, Vol. 45, No. 1, pp. 5-32. (in English)

- Bae, S.W. (2019), *Forecasting Property Prices Using the Machine Learning Methods: Model Comparisons*, Ph.D. dissertation, Dankuk University, Seoul, Korea, 177p.
- David, Rohde., Jonathan, Corcoran. and Prem, Chhetri. (2010), Spatial Forecasting of Residential Urban Fires: A Bayesian Approach, *Environment and Urban Systems*, Vol. 34, pp. 58–69. (in English)
- Hong, S.Y., Cho, S.H., Kim, M.S. and Moon, I. (2019), Fire Prediction based on Weather and Fire Data using Artificial Neural Network, *Fire Science and Engineering*, Vol. 19, No. 7, pp. 275-281. (in Korean with English abstract)
- Hwang, W.U., Go, M.H. and Yang, B.S. (2004), Cavitation Condition Monitoring of Butterfly Valve Using Support Vector Machine, *Transactions of the Korean society for noise and vibration engineering*, Vol. 14, No. 2, pp. 119-127. (in Korean with English abstract)
- Hastie, C. and Searle, R. (2016), Socio-Economic and Demographic Predictors of Accidental Dwelling Fire Rates, *Fire Safety Journal*, Vol. 84, pp. 50-56. (in English)
- Jang, H.S. (2018), *Predictive evaluation of fecal E. coli occurrence using machine learning*, Ph.D. dissertation, Kunsan National University, Jeonbuk, Korea, 121p.
- Kim, M.J. (2014), *A Study on Fire Prediction Model Development Using Data Mining*, Master's thesis, Kangwon University, Kangwon, Korea, 52p.
- Kim, S.J. and Ahn, H.C. (2016), Application of Random Forests to Corporate Credit Rating Prediction, *Journal of Industrial Innovation Research*, Vol. 32, No. 1, pp. 187-211. (in Korean with English abstract)
- Kim, H.R., Sin, J.W., Park, Y.J., Lee, H.P. and Moon, K.A. (2010), A Study on the Statistical Analysis of Fire Patterns in Gangwon Province, *Fire Science and Engineering, Proceedings of the Korea Institute of Fire Science and Engineering Conference*, 23 April, Kyungmin University, Korea, pp. 419-423.
- Lee, C.L. (2015), *Estimating Single-family House Prices Using Non-parametric spatial Models and an Ensemble Learning Approach*, Ph.D. dissertation, Seoul National University, Seoul, Korea, 165p.
- Lee, H.L. (2020), *Real-time Prediction of Penetration Rate of Shield TBM based on Machine Learning*, Ph.D. dissertation, Inha University, Incheon, Korea, 132p.
- Lee, C.L. and Park, K.H. (2016), Application of machine learning models for estimating house price, *Journal of the Korean Geographical Society*, Vol. 51, No. 2, pp. 219-333. (in Korean with English abstract)
- Lee, H.P., Lee, S.C., Hwang, M.J., Park, Y.J., Moon, K.A. and Kim, H.B. (2010), A Study on the Analysis of Fire Patterns using the Decision Tree Analysis Method, *Fire Science and Engineering, Proceedings of the Korea Institute of Fire Science and Engineering Conference*, 30 October, Seoul, Korea, pp. 349-353.
- Lu, S., Liang, C., Song, W. and Zhang, H. (2013), Frequency-size Distribution and Time-scaling Property of High-casualty Fires in China: Analysis and Comparison, *Safety Science*, Vol 5, pp. 209-216. (in English)
- National Fire Agency. (2019), *National Fire Data System*, <https://www.nfds.go.kr/>, 28 June 2021.
- Park, K.H. (2018), *Bagging, Random Forest and Ensemble SVM Comparison studies*, Master thesis, Inha University, Incheon, Korea, 25p.
- Rifan, Ardianto1. and Prem, Chhetri1. (2019), Modeling Spatial–Temporal Dynamics of Urban Residential Fire Risk Using a Markov Chain Technique, *Int J Disaster Risk Sci*, Vol. 10, pp. 57–73. (in English)
- Seo, J.D. (2016), Foreign Exchange Rate Forecasting Using the GARCH extended Random Forest Model, *Journal of Industrial Economics and Business*, Vol. 29, No. 5, pp. 1607-1628. (in Korean with English abstract)
- Seo, M.S. and Yoo, H.H. (2020) Significance Analysis of Facility Fires Though Spatial Econometrics Assessment, *Journal of the Korean Society of Surveying, Geodesy, Photogrammetry, and Cartography*, Vol. 38, No. 3, pp. 281-293. (in Korean with English abstract)
- Vapnik, V. and Chervonenkis, A. (1964), A Note on One Class of Perceptrons, *Automation and Remote Control*, Vol. 25, No. 1, pp. 103-109. (in English)
- Vapnik, V. and Lerner, A. (1963), Pattern Recognition Using Generalized Portrait Method, *Automation and Remote Control*, Vol. 24, No. 6, pp. 774-780. (in English)